Statistics and the Environment

George E. P. Box[1]

Department of Statistics, University of Wisconsin, Madison 53715

The Problem

It seems only a little time ago that we were concerned with matters which now seem comparatively trivial. We had for some time lived with the knowledge that our survival was threatened by nuclear attack by a foreign enemy, but it seems only recently that we have noticed a more insidious threat of our own making. Most of us now recognize that we are well on the way to destroying ourselves by overpopulation, pollution, the frittering away of our raw materials, and the poisoning of our food by inadequately tested chemicals.

Opinions differ as to how long it will take before various predictable crises occur and how much each problem will complicate the solution of the others, but it is very clear that we will be hard pressed and we will be lucky to escape by the skin of our teeth. The truth is that although we are called on to meet very difficult problems of great urgency we know pathetically little of the facts. So we must learn fast.

Now it is precisely this ability to learn fast that has got us into our present difficulties. It was only a few hundred years ago that men's minds seriously turned to the question of how the, normally very slow, process of learning by chance experience might be accelerated. Scientific method, the secret of learning fast, has altered the normal birth and death process, yielding perhaps a more comfortable world but at the cost of world overpopulation. Scientific method has provided us with motor cars and factories producing convenient products, but the by-products of both are threatening the air we breathe and the water we drink. Furthermore, their insatiable appetite for raw materials is stripping the earth of its irreplaceable treasure. Scientific method has provided us with conveniently packaged foods with chemical additives which make them taste good, look good, and last a long time on the shelves of the supermarket, but pharmacologists will tell you that it is almost impossible to keep up with the flood of these new substances which we ingest, and to be sure what are their long term effects on human beings.

So we are hopeful that the same scientific method which has in a period of a few hundred years got us to where we are now, can in a few decades get us to where we would like to be.

I believe it can, but with two provisos:

First, we must release, by public education, the will to make it happen.

Second, because with so little time we cannot afford inefficient investigation, we must catalyze the learning process still further. The catalyst is the proper use of Statistical Methods.

Science and Statistics

It was Lord Kelvin who said, ''When you can measure what you are speaking about and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.'' But, in case that should seem too much an encouragement to those who believe that mere unthinking accumulation of numbers is synonymous with good science

[1] Supported by the United States Office of Army Research under grant number DA-ARU-D-31-124-72-G162.

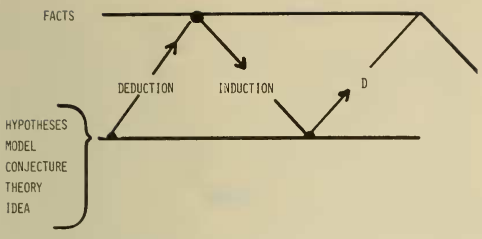

Fig. 1. The iterative learning process.

and will of itself solve the problem, I hasten to add the well known words attributed, among others, to Mark Twain, a contemporary of the noble lord's: "There are three kinds of lies—lies, damn lies, and Statistics."

What then is scientific method and what part does Statistics play within it?

Scientific method is a process of controlled learning. The object of statistical method is to make that learning process as efficient as possible.

Learning is an iterative process, illustrated in Fig. 1, in which a hypothesis (or theory or model or conjecture) leads by a process of deduction to certain consequences which may be compared with known facts. Usually the consequences and the facts fail to agree, leading by a process called induction to modification of the hypothesis. Thus a second iteration is initiated, the consequences of the modified hypothesis are worked out and again compared with facts (old or newly acquired) which, with luck, leads to further modification and to further gaining of knowledge.

This process of learning can be thought of in terms of the feedback loop shown in Fig. 2, where discrepancy between the facts and the consequences of the initial hypothesis H leads to the modified hypothesis H′. This view makes it clear why there is no place in science for the man who wants to demonstrate that he has always been right. For it is by arranging matters so that there is maximum opportunity to find out where he may be wrong, that most progress is made.

Suppose at a certain stage in an investigation the situation is that shown in the bottom half of Fig. 3. A hypothesis H concerning the state of nature has been formulated, leading to certain consequences that have been compared with the facts deduced from analysis of the available data. Discrepancies have suggested a modification from H to H′. Consequences of H′ may now be in accord with the data analysis or may still be discordant. When it is not clear what modification should be made to an unsatisfactory hypothesis or, alternatively, when confirmation of an apparently satisfactory hypothesis is needed, further data must be sought. Depending on the context, the further data may come from a designed experiment, a sample survey, or already existing results. Whatever the source of the data, careful attention must be given to its selection or design. As illustrated in Fig. 3, the direction of the effort at data getting will inevitably depend on our latest view of the state of nature and the hopes and fears which surround that view.

While, at a particular stage, the conjectured state of nature may be false or at least inexact, the data themselves are generated by the true state of nature. It is because of this that the comparison of

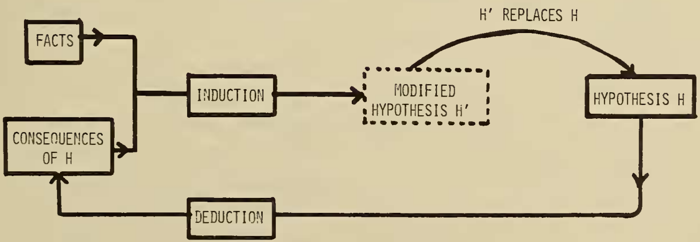

Fig. 2. The learning process as a feedback loop.

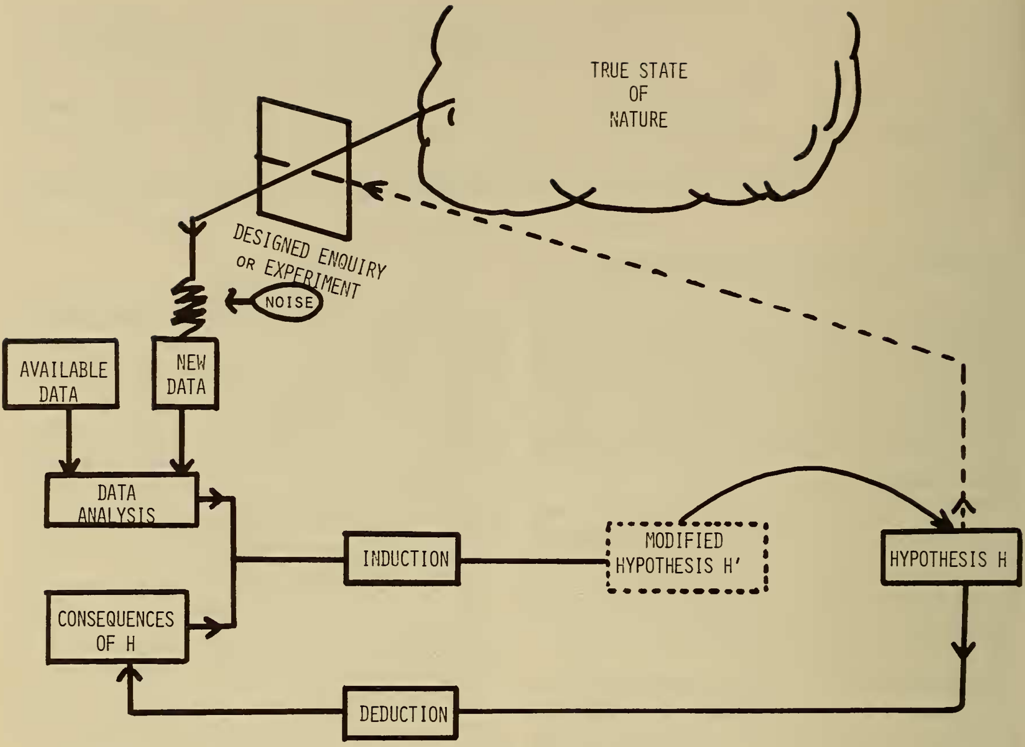

Fig. 3. Data analysis and data getting in the process of scientific investigation.

successively conjectured states of nature with actual data can lead to convergence on the truth. Even if we could see such data free of experimental error, however, the task of discovery would usually not be easy because of the complexity of the systems that need to be studied. So in practice in addition to complexity, we have to cope with an added difficulty—that the data contains experimental error (or noise), which tends to mislead.

Scientific investigation, then, is not easy, and obviously the process we have described depends crucially on the scientific wit and subject matter knowledge of the investigator. The statistician's job is to advise and assist the investigator in two crucial tasks, so as to allow the investigator to employ his talents most efficiently. These tasks are:

(1) deciding what would be appropriate data to get at each stage of the investigation. Broadly we can call this the *design* problem.

(2) deciding what the data entitles us to believe at each stage of the investigation. We can call this the *analysis* problem.

Of the two, *design*—the decision as to what are the appropriate data to get—is of paramount importance. This is equally true whether by actual design of an appropriate experiment, the planning of a suitable sample survey, or the proper choice of a data base. No amount of skill in data analysis can extract information which is not there to begin with. The second task of the statistician, although not so vital as the first, is still very important. Inappropriate analysis of data can produce unjustifiable conclusions or fail to discover justifiable ones. *Worse,* it can fail to unearth those hints of, perhaps unexpected, phenomena which often catalyze the investigator's progress to a solution. In any case, inappropriate analysis of data will greatly hamper convergence of the scientific iteration.

In summary then, we learn through numbers. But what numbers or data should we try to get and what do they mean when we have them? These are the questions that good statisticians are trained to answer. It is very easy to acquire useless or irrelevant data. It is very easy to be misled by data once they are acquired. The design of *each stage* of an enquiry so as to produce useful data with the minimum of time and expense, and the analysis of data of each stage so as to produce, not only valid conclusions, but also valuable hints on how the investigation ought to proceed, these are the two critical tasks in which the statistician plays a key role.

Part of the statistician's job is also, I think, to encourage and accompany the scientist in the slightly schizophrenic role that he has perforce to play.

Having entertained a tentative model (hypothesis, etc.) it is up to the statistician to see that fully efficient means are used to investigate the consequences of that model. That is the inference step in Fig. 4. However, having then produced the best analysis possible, supposing the model to be accurate, he must now change his stance from that of a sponsor to a critic. He becomes a doubting Thomas prepared to find fault by inspecting residuals for suspicious features, etc. This criticism can lead to modification of the model, either at once or at some time after more data has been taken.

Switching alternately from sponsor to critic and back again is a painful business but one which we must steel ourselves to pursue. The Pygmalions who have fallen in love with their models somewhere along the way are a nuisance and a hindrance to progress.

Another part of the statistician's job is to make sure that Statistics and Computers do not separate the investigator from his data but, on the contrary, help him to see his data from many different angles. We must remember that the best induction machine so far devised is the human mind, and if modern methods of dealing with data result in separating the investigator from his data, they are almost certainly doing more harm than good.

Going now into a little more detail, what then are some of the difficulties that appropriate use of statistical methods can alleviate or avoid?

Coping with Natural Variation

We live in a world which is universally variable. How much air a man breathes depends on the particular man, his temporary physiological state, the atmosphere he is presently in, and so forth. And yet, until quite recently, attempts were made to study variable phenomena in an entirely deterministic manner. Variation was frowned upon, as if disapproval

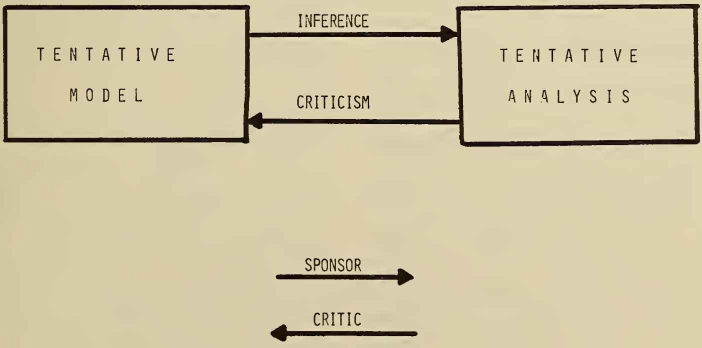

Fig. 4. Statistical analysis as an iterative process.

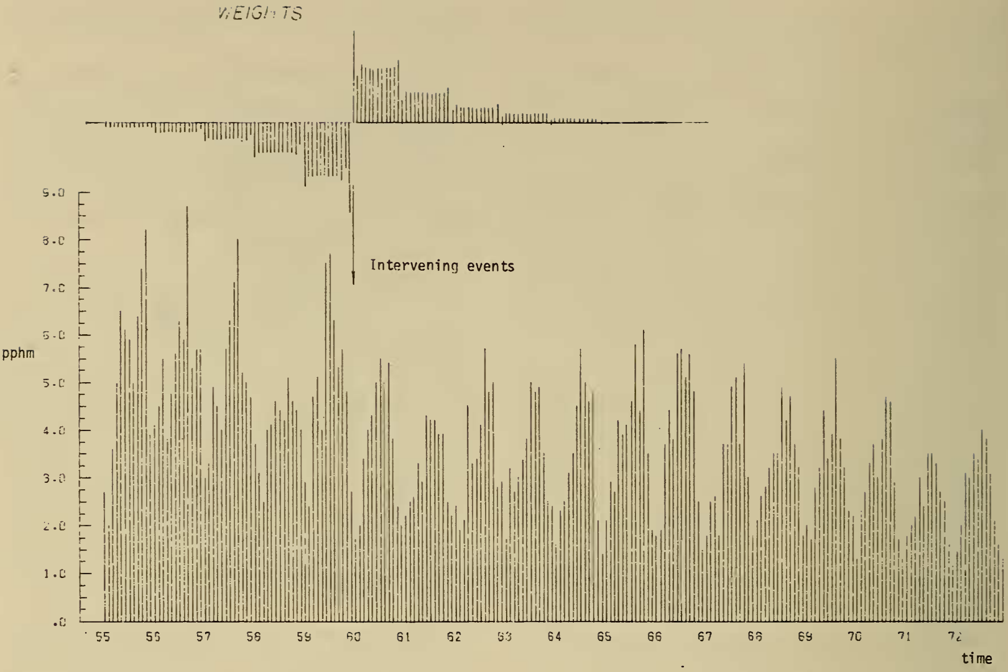

Fig. 5. Monthly average of hourly readings of O_3 (pphm) at downtown Los Angeles (1955–1972), with the weight function for estimating the effect of intervening events in 1960.

would make it go away, and probability statements were treated as in some way unsatisfactory. There was little readiness to admit that everything varies and, except perhaps from God himself, every statement, if exactly made, would have to be a probability statement.

Increasing Accuracy by Exploiting the Variational Structure

Environmental data are usually highly variable. It is by facing this fact, rather than running away from it, that we can solve some of our problems. Indeed, it is a fascinating fact, that it is the *structure* of the variation or noise, which determines how we can extract the information which the data contain. As an illustration, Fig. 5 shows monthly averages of oxidant (O_3) levels observed in downtown Los Angeles from 1955 to 1972. These data are highly seasonal and variable. About the beginning of 1960 two events occurred which might have been expected to change these levels. These events were the diversion of traffic by the opening of the Golden State Freeway and the coming into effect of a new law (Rule 63), which reduced the allowable proportion of reactive hydrocarbons in the gasoline sold locally. By a study of the structure of the variation it is possible to obtain (Box and Jenkins, 1970; Box and Tiao, 1965, 1973; Tiao et al., 1973) a valid and most sensitive test of the possibility that the events in January 1960 changed the oxidant level and to estimate the change. For this data the estimate turns out to be −1.10 ± 0.10 p.p.h.m. The function shown at the top of the diagram displays the manner in which the data are weighted in the optimal difference estimate. As common sense might expect (i) most weight is given to values obtained immediately before and after the events and remote data are suitably discounted, (ii) the weighting is automatically chosen so that seasonal effects are eliminated.

I believe that the use of ''Intervention

Analysis" such as the above, in which difference equation models are used to represent dynamic and stochastic systems, has much to contribute in uncovering possible effects of public policy changes. For example, it could show the effect of the opening of a nuclear power station on the ecology of the river from which cooling water is drawn and returned. It is clear that studies of this kind are vital to the intelligent framing of new laws.

We owe to Sir Ronald Fisher the concept that we can exploit the patterns of natural variation in data to *design* enquiries and experiments so that errors are minimized. For example, randomized block designs and stratified sampling plans can eliminate major sources of disturbance and ensure that important comparisons are made within the least variable material.

Another tool which should, I believe, find much application is the use of components of variance to improve tests of environmental quality. The analysis of variance table used in the analysis of data from the randomized block designs I mentioned above may also be employed in conjunction with suitable designs to estimate components of variation, for example, in tests of environmental quality. Suppose we take a sample from a stream and perform an analysis. How accurate is the result we get? What do we mean by that question? Certainly not how closely repeated chemical analyses of that same sample would agree with one another. What we want to know is how nearly does our analysis give a picture of the quality of that stream at that time and place.

An appropriate study of components of variance—how much variation is associated with chemical analysis, how much with the sampling method, how much with change of location in the river, together with knowledge of how much it will cost to take a sample and perform a chemical analysis—enables us to devise a testing scheme which can be dramatically more accurate and economical than one naively chosen.

Causation and Correlation

Many years ago when I studied statistics at University College London there was a plot of some data which none who saw could easily forget. On the x axis was the number of storks' nests observed each year in a certain town; on the y axis was the corresponding human birth rate for that year. The data showed an almost perfect straight line relationship. It is perhaps superfluous to explain that the correlation arose because, over the period of years in which the data were taken, the stork population was increasing and so was the human population. It is also unnecessary to point out that our over-population problem will not be solved by shooting storks.

In case these remarks should seem frivolous we must remember that it was precisely this kind of question which was debated in some of the early discussions on smoking and lung cancer and which bedevil much data analysis in other fields.

Again it was Fisher who showed how in planned experimentation the introduction of randomization could break the purely correlative chain and enable causation to be distinguished. In cases where planned experiments are not possible the situation is always very tricky, and very careful analysis is needed to decide in any given case precisely what the data allow us to conclude.

Complexity in the Face of High Noise Levels

Many of the phenomena we face in considering the environment are complex. To cope with problems which are complex, as well as being obscured by experimental error, we would be wise to welcome whatever help we can get. Even though the complexity of problems is admitted, the idea that variables should only be studied one at a time dies hard. The one variable at a time method would, of course, only be a satisfactory mode of study if nature were so obliging as to have its variables affect the environment independently. Again, it was Fisher who pointed out that by the use of suitable design the effect of experimental error could be averaged out at the same time that provision was made for the estimation of complex effects. Designs of this kind may be used, not only for empirical descriptions of phenomena, but also for testing mechanisms. This is done by treating as data the estimated constants of the system. If the model is correct, these should remain constant when extraneous conditions are varied. When, as is usually the case initially, the model is not wholly correct, analysis of the changes in the ''constants'' provides a valuable diagnostic tool for model testing, pointing to where the model needs attention. Endelman (1973) has recently used these methods at Wisconsin to study nitrogen changes in the soil and soil water. In many ways this study was a model one, in which the Departments of Soil Science, Chemical Engineering, and Statistics all cooperated.

While on the subject of complexity a word should be said about the models needed to represent complex phenomena. In any given investigation it seems to me we can err in two ways. We can have too simple a model or too elaborate a model. My recent experience has been that investigators have often erred in building models that are too elaborate. There is a tendency to try to model each step that the investigator can imagine, whether there is strong evidence that that step really occurs in the system or not, whether the step affects the solution or not, and whether the data could possibly supply any information about that step or not. Even if he had a 50% chance of being right about any given step, the investigator need only introduce a few such steps into a system and the chance of error becomes overwhelming. My experience is that we must borrow William of Occam's razor and use it rather ruthlessly to remove deadwood. Usually, models are best built up from simple beginnings, elaboration being introduced only as it is shown to be necessary by actual comparison with data, as in Fig. 3.

The Peril of the Open Loop

Perhaps of all the problems that face us, whether personal, professional, scientific or statistical, the most menacing of all is the danger of the open loop.

I have spoken of the process of scientific learning in terms of a feedback loop. If the loop is open, learning stops, of course. The idea applies more generally. As an earlier speaker has so ably pointed out, feedback is essential between scientists and legislators; otherwise, even when the scientists know what to do, it cannot get done.

As another example, I recently attended a seminar where the speaker was building a pollution model for a city. The method he used was to calculate by dead reckoning the amount of every substance going into the atmosphere over each small area of the city. For example, he could calculate over, say, a given hundred yards square area, how much rubber was worn off the tires of automobiles passing through that area and hence presumably going into the atmosphere. There was nothing wrong with that, but I was surprised to hear him explain, as he commenced his seminar, that there were two kinds of modelling—his kind based on dead reckoning and statistical modelling based on data. Learning happens surely only when the loop is closed and what can be calculated from dead reckoning is compared with what the data actually say.

The Supply of Competent Statisticians

Perhaps finally I should say something about the supply of statisticians. A little while ago I saw a report prepared by a distinguished panel of mathematicians on the current need for graduate training in mathematics and mathematically related subjects. One conclusion was that since a principal outlet for Ph.D. mathematicians was as university teachers and since the great expansion of the universities had now ceased, we must plan for a major cut-back in the production of Ph.D.'s or face the possibility of producing a glut of unemployed mathematicians. I was alarmed because the "mathematically related subjects" which the report claimed to cover included statistics!

Now whatever may be true about the future need for pure mathematicians, the fact is that we face a scarcity of trained statisticians competent to deal with real problems. Furthermore as, one by one, the various environmental crises become more obviously imminent and the need for hard facts on which to take sensible action becomes inescapable, the demand for such people will markedly increase. It takes many years to produce a properly trained statistician. It cannot be over-emphasized that steps must be taken now not to restrict but to expand the educational facilities available for the training of competent statisticians.

How do we get competent statisticians? Neither surely by producing mere theorem provers nor mere users of a cook book. A proper balance of theory and practice is needed and, most important, statisticians must learn how to be good scientists, a talent which, I think, has to be learned by example. At Wisconsin, we have taken a number of steps to help this along:

- To obtain any graduate degree, a student must have spent a period of time in the Statistical Consulting Lab working with the statistician in residence and other faculty to deal with clients' problems. This counts as a course for credit, and no student can graduate without passing this course.
- The Masters Degree, which all students are encouraged to take, whether or not they proceed to a Ph.D., is not a "failed Ph.D." degree but is awarded on their demonstrated competence to becoming a practicing statistician.
- A Monday night beer session is held in the basement of my house where research problems are discussed on an ongoing basis.
- The department is deliberately diversified with joint appointments and research interests in engineering, business, medicine and agriculture.
- Students act as research assistants in projects such as the Analysis of the Los Angeles Air Pollution data, the improvement of operating methods for the local sewage works, etc.

When we look at the history of the subject of statistics itself, there is no doubt that it develops most rapidly when there is active feedback, with practical problems initiating new theory and new theory in turn showing new ways to handle real situations. I believe we are moving now into a period of great statistical activity where, because of the service it will render to the community, our science will come into its own. In doing so, it will inevitably undergo new and exciting development.

Literature References

Box, G. E. P., and Jenkins, G. M. 1970. *Time Series Analysis: Forecasting and Control.* Holden-Day.

Box, G. E. P., and Tiao, G. C. 1965. A change in level of a non-stationary time series. *Biometrika,* Vol. 52.

Box, G. E. P., and Tiao, G. C. 1973. "Intervention Analysis with Applications to Environmental Problems", Technical Report No. 335, Department of Statistics, University of Wisconsin, Madison.

Endelman, F. 1973. "Systems Studying of the Transport and Transformations of Soil Nitrogen", Ph.D. Thesis, University of Wisconsin, Madison.

Tiao, G. C., Box, G. E. P., and Hamming, W. J. 1973. "Analysis of Los Angeles Photochemical Smog Data: A Statistical Overview", Technical Report No. 331, Department of Statistics, University of Wisconsin, April.